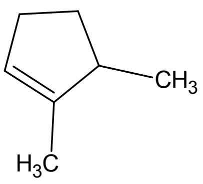
The IUPAC International Chemical Identifier (InChI)

The question of clearly identifying a chemical has been present almost since the beginning of modern chemistry. Two-dimensional depictions of chemical structure have long been the most popular means of expressing the identity of a compound whose bonding was known. However, in writing and speaking, a pronounceable name was often required, leading to the development of a variety of systematic naming methods. The most widely adopted of these is a rule-based system of the International Union of Pure and Applied Chemistry (IUPAC). NIST has worked jointly with IUPAC to create a uniform and open standard that could be adopted by the entire chemical community

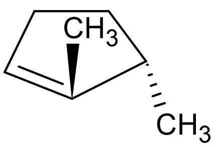
D.Tchekhovskoi, S. Stein (Div 838)
S. Heller (Guest Researcher)

For many chemicals, names can be difficult to generate or interpret, and so co-called "trivial" names are still often used for common chemicals, as are complex structural drawings. A structural depiction is closer to the true meaning of chemical identity and can allow a rapid understanding of the properties of the chemical that a long text name can never provide. However, phenomena of tautomerism (rapid hydrogen atom migration) and (de)protonation change chemical structure thus hiding the chemical identity of compounds. In addition, the same chemical structure may be drawn in different ways, making it hard to establish equivalence of the drawings.

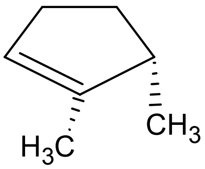
Representation problems grew with the increasing use of digital communication in chemistry, since graphical pictures of chemical structures could not be readily interpreted by computers. This led to a joint NIST/IUPAC project to create a uniform and open standard that could be adopted by the entire chemical community – The IUPAC International Chemical Identifier (InChI). The goal was to create a flexible and error-tolerant naming system that would allow computers to uniquely identify a chemical, regardless of how it is drawn and based entirely on the connectivity of the molecule – that is, what atoms are connected to what other atoms. To accomplish this a lot of what is normally viewed as "chemical information" was discarded and the molecules were trimmed to the minimum information needed to differentiate one from the other. In addition, a layered approach was developed to deal with some of the more complex issues of chemical structure. For example, the two molecules in the figure differ only in that one has the two methyl groups on the same side relative to the plane of the ring (on the right – called cis) and the other has the two methyl groups on opposite sides of the ring (in the center – called trans). On the left is a diagram that can be used to represent either of the molecules. The left diagram shows only the connectivity and does not specify if the molecule is cis or trans. The problem encountered prior to InChI is that the data retrieval was often dependent upon the way the molecule had been drawn. There is often a need to distinguish between the cis and trans form, and often a need to search for all possible forms, including cases where the configuration of the molecule was not known or it was known that a mixture was present.



unspecified



trans



cis

The approach taken in developing InChI is a layered approach. This allowed as much information as was known to be specified, allowing two representations of the same substance given at different degrees of precision to be matched. The most common deficiencies in structural representation results from such issues, and the layering allows for natural means of resolving these problems.

The IUPAC International Chemical Identifier (InChI) has been released by IUPAC in April 2005.

The identifier project is an example of open source software. In addition, a number of tools allow users to easily analyze their own structures using the freely available compiled code. As InChI becomes more widely adopted, it is expected that it will enable a standardized referencing and search for chemical structures both over the Internet and in proprietary databases.

The IUPAC InChI has been adopted/used by a variety of entities:
PubChem, a major resource for medical research at the National Institutes of Health (NIH), has adopted InChI as a standard for identifying and searching for compounds. For example, see:
http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=6986.
InChI has been integrated by ACD Labs* in their widely used commercial drawing program, ChemSketch, as well

as its freeware version that is distributed for home and student use, see:
http://www.acdlabs.com/download/chemsk.html.
The Chemistry WebBook had adopted InChI and displays the identifier for all data in the WebBook thus making it possible for the very large audience that uses the WebBook to gain access to the identifier, see for example: compare the structures and representations of 2-hexene.
The Environmental Protection Agency has adopted InChI for use in their Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network, see:
http://www.epa.gov/nheerl/dsstox/MoreonINChI.html
The Web of Science, (Thomson Scientific) one of the most widely used information sources for searching the scientific literature, has adopted the identifier.

There are many other uses of the identifier – and since the software is distributed free of charge we do not even know if we are aware of all of them. One example is the Compendium of Pesticide Common Names (Alan Wood, UK), a site that allows users to find the common name as well as the structure for a pesticide, see: http://www.hclrss.demon.co.uk/. The wide number of adoptions has been made easy by a range of software solutions developed as a part of the project. Many of these are available from IUPAC.

Future Plans:

To create the needed protocol (and the associated software) to allow users to verify InChI string created anywhere. This would ensure that any code that was ported to other operating platforms would still generate valid InChI strings.

In addition, we are seeking the best way to extend InChI to include polymers, phase, and excited states.

Publications and Presentations:

- Chemical 'Naming' Method Unveiled, *Chem. & Eng. News*, 22 Aug 2005 Analysis of a Set of 2.6 Million Unique Compounds gathered from the Libraries of 32 Chemical Providers, A. Monge, A. Arrault, C. Marot and L. Morin-Allory, presented at the *10th Electronic Computational Chemistry Conference*, April 2005
- International chemical identifier goes online, *Chem. World*, 16 May 2005
- Application of InChI to Curate, Index, and Query 3-D Structures, M.D. Prasanna, J. Vondrasek, A. Wlodawer and T.N. Bhat, *Proteins: Structure, Function, and Bioinformatics*, 2005, **60**, 1-4
- Enhancement of the chemical semantic web through the use of InChI identifiers, S.J. Coles, N.E. Day, P. Murray-Rust, H.S. Rzepa and Y. Zhang, *Org. Biomol. Chem*., 2005, **3**(10), 1832-1834
- InChI FAQ, by Nick Day (Unilever Centre for Molecular Informatics, Cambridge University)
- Representation and Use of Chemistry in the Global Electronic Age, P. Murray-Rust, H.S. Rzepa, S.M. Tyrrell and Y. Zhang, *Org. Biomol. Chem.*, 2004, 3192-3203 [www.ch.ic.ac.uk/rzepa/obc/]
- That InChI feeling, *Reactive Reports*, issue 40, Sep 2004
- Unique labels for compounds, *Chem. & Eng. News*, 2 Dec 2002
- Chemists synthesize a single naming system, *Nature*, 23 May 2002
- That InChI feeling ... *The Alchemist*, 24 Apr 2002
- What's in a Name? *The Alchemist*, 21 Mar 2002

**Disclaimer:*

Certain commercial equipment, instruments, or materials are identified in this document. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.